DOCUMENT RESUME

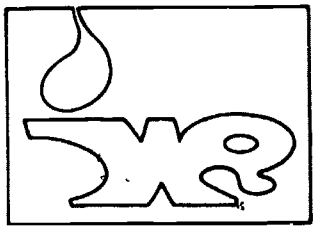
ED 219 246 SE 038 744

AUTHOR Romberg, Thomas A.
TITLE The Development and Validation of a Set of Mathematical Problem-Solving Superitems. Executive Summary of the NIE/ECS Item Development Project.
INSTITUTION Wisconsin Center for Education Research, Madison.
SPONS AGENCY Education Commission of the States, Denver, Colo.; National Inst. of Education (ED), Washington, DC.
PUB DATE Jan 82
CONTRACT 02-81-20321
NOTE 28p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Research; Elementary Secondary Education; Evaluation; Evaluation Methods; *Mathematics Achievement; *Mathematics Education; *Problem Solving; *Test Construction; *Testing
IDENTIFIERS *Mathematics Education Research

ABSTRACT

This document describes seven steps that were followed to develop and validate a pool of mathematical problem-solving situations and a set of questions for each situation which were designed to provide information about students' qualitatively different levels of reasoning ability. For each stage a description is presented of the work that was carried out and what was accomplished. It is noted that a strategy of developing a set of "structured super-items" was followed for each of a set of problem-solving situations. It is concluded from this effort that a content-valid set of superitems was successfully constructed for administration. Further, construct validity of the items was established in relationship to an underlying theory of response outcomes, and the utility of the superitems was noted as demonstrated. It is felt that since the goals of the study were obtained, the way to a more useful set of items which could be used in large scale assessment projects has been pointed out. (MP)



EXECUTIVE SUMMARY OF THE NIE/ECS ITEM DEVELOPMENT PROJECT

The Development and Validation of a Set of Mathematical Problem-Solving Superitems

by Thomas A. Romberg

January 1982

Wisconsin Center for Education Research
an institute for the study of diversity in schooling

Executive Summary of the NIE/ECS Item Development Project

THE DEVELOPMENT AND VALIDATION OF A SET OF

OF MATHEMATICAL PROBLEM-SOLVING SUPERITEMS

by Thomas A. Romberg

Thomas A. Romberg
Principal Investigator

Wisconsin Center for Education Research
The University of Wisconsin
Madison, Wisconsin

January 1982

This material is based upon work supported by the National Institute of Education and the Education Commission of the States under Contract No. 02-81-20321 with the Wisconsin Center for Education Research. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of ECS, NIE, or the Department of Education.

TABLE OF CONTENTS

Abstract

This is a summary report which describes the seven steps that were followed to develop and validate a pool of mathematical problem-solving situations and a set of questions for each situation which were designed to provide information about students' qualitatively different levels of reasoning ability.

For each stage a description is presented of the work that was carried out and what was accomplished. From this effort we have concluded that we were able to construct a content valid set of superitems for administration, to establish the construct validity of the superitems in relationship to an underlying theory of response outcomes, and to demonstrate the utility of the superitems.

Introduction

The purpose of this summary report is to describe the steps that were followed to develop and validate a pool of mathematical problem-solving situations and a set of items for each situation which were designed to provide information about students' qualitatively different levels of reasoning ability.

The strategy followed was to develop a set of "structured superitems" for each of a set of problem-solving situations. The method for creating a pool of situations and questions was based on Cureton's (1965) notion of "superitems" (a set of test questions based on a common situation or stem). The structure for the superitems was based on Collis and Biggs' SOLO taxonomy used to classify the structure of observed learning outcomes. The superitems were prepared to be administered to students of 9, 11, 13, and 17 years of age. The superitems then were administered to over 300 students at each age level to examine both their validity and the utility of the procedure for large scale assessments. Since the goals of this study were attained, we believe a more useful assessment procedure for this critical aspect of mathematics can be used for large scale assessments.

The project was funded by the Education Commission of the States (with funds supplied by the National Institute of Education). The resulting items could be useful in future National Assessment of Education Progress (NAEP) studies in mathematics.

To accomplish the goals of this study, a seven-stage project was designed.

Stage 1. December to March 1981--Problem Situation Development.

For the student populations, a set of problem situations was developed.

Stage 2. March to May 1981--Basic Validity Check.

Each problem situation was examined by classroom teachers at the respective grade levels to check on the appropriateness of the concepts and prerequisite skills for students of those ages.

Stage 3. April to July 1981--Superitem Development.

At this stage, sets of items for each situation were written, reviewed, and tried out with a small sample of students under the direction of Professor Kevin Collis. The items were again reviewed by graduate students to check the items for their mathematical appropriateness and their fit to the SOLO taxonomy. This tryout was done to ensure that students could read the items and follow directions and to see if there were any procedural problems.

Stage 4. July to September 1981--Preparation of Trial Materials.

At this stage, the set of situations and superitems appropriate for the target population was organized into batteries for administration to a large population of students.

Stage 5. September 1981--Administration of Batteries.

Early in the school year the batteries were administered to a population of students.

Stage 6. October through December 1981--Data Analysis.

All test booklets and questionnaires were scored and analysis of the data was carried out at this stage.

Stage 7. December through January 1982--Report Preparation.

Summary of Results at Each Stage

Stage 1

Initially 40 problem stems were written for six content categories: numbers and numeration; variables and relationships; size, shape, and position; measurement; statistics and probability; and unfamiliar. These categories correspond to the five NAEP content designations and an additional area termed "unfamiliar." Then for each item stem, three to five questions were written which reflected comprehension, application, and analysis objective categories previously used by Wearne and Romberg (1977).

Stage 2

Twenty classroom teachers (8 twelfth-grade teachers, 6 seventh-grade teachers, and 6 fourth-grade teachers) were recruited to judge the superitems on three dimensions. The dimensions teachers were to consider were content, whether the item stem fit the six content categories; reasoning levels, whether each question in a superitem fit one of three objective categories; and appropriateness, whether the questions in each superitem were appropriate for students at the teacher's grade level.

With the exception of the seven "unfamiliar" stems, content agreement by the teachers appear to be fairly consistent with the content categories for which the items were written. Overall agreement of teacher judgments with the intended cognitive level for each question was good. Finally, 74.5% of the questions were considered appropriate. However, the judgments by teachers at different grades were considerably different. Almost all

of the questions were considered appropriate by the twelfth-grade teachers while only 49% of the questions were considered appropriate by fourth-grade teachers. Many questions were considered either to be too difficult for fourth-grade children or on content they had not covered.

Stage 3

Beginning April 1981 when Professor Collis arrived in the U.S., the questions for each item were rewritten according to the SOLO taxonomy (Collis & Biggs, 1979). The taxonomy was designed as a response model, the basic idea being that the child is given information or data and asked a question which can be answered by reference to that information. The child's response is classified as belonging to one of five levels according to the way in which the response is structured.

For this project we hypothesized that by using the SOLO framework one could develop a series of questions based on the stem that would require a more and more sophisticated use of the information from the stem in order to obtain a correct result. This increase in sophistication should parallel the increasing complexity of structure noted in the SOLO categories.

The criteria we used to write questions so that a correct response to each question would be indicative of an ability to respond to the information in the stem at least at the level reflected in the SOLO structure of the particular question were:

Pre-structural (P)	Use of no information from the stem or no response.
Uni-structural (U)	Use of one obvious piece of information coming directly from the stem.

Multi-structural (M)	Use of two or more discrete closures directly related to separate pieces of information contained in the stem.
Relational (R)	Use of two or more closures directly related to an integrated understanding of the information in the stem.
Extended Abstract (E)	Use of an abstract general principle or hypothesis which is derived from or suggested by the information in the stem.

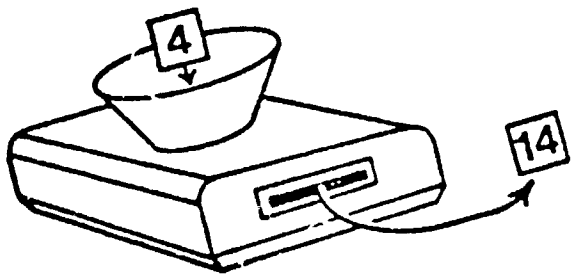
An example of items constructed in this manner is shown in Figure 1. The stem provides information and each question that follows requires the student to reason at a different level in order to produce a correct response.

Selected items were administered to children from Shawno elementary and middle schools, from Cottage Grove elementary school, and from Monona Grove High School.

Results indicated a great deal of consistency in the SOLO levels recorded for each child and also for children at the same grade level. Variance in levels for each child was almost wholly within one response category of the level of reasoning generally observed for the grade.

Based on this information, all superitems were reviewed and many revisions were made. At this stage then, in June 1981, six graduate students in mathematics education at the University of Wisconsin-Madison responded to the pool of 40 superitems. The graduate students were instructed to work each item and classify each as being primarily in one of the six content categories. In addition, the students were to identify, for each question in the items, the level of reasoning likely to be employed.

This is a machine that changes numbers. It adds the number you put in three times and then adds 2 more. So, if you put in 4, it puts out 14.



U. If 14 is put out, what number was put in?

M. If we put in a 5, what number will the machine put out?

R. If we got out a 41, what number was put in?

E. If x is the number that comes out of the machine, when the number y is put in, write down a formula which will give us the value of y whatever the value of x.

Figure 1. Example of a superitem written to reflect the SOLO taxonomy.

The results indicated a generally high level of agreement for both content and level of reasoning categorizations. Again, only the index of agreement for the "unfamiliar" stems was particularly low. Thus, since the indices of agreement were high for judgments about content and particularly for judgments on level of reasoning, we felt content or face validity of the superitems has been demonstrated.

Stage 4

At this time a final technical review of all items was carried out. This review was in part editorial; for example, wording was simplified, tenses were checked, and agreement in terminology and symbols among the stem and all questions for each item was inspected. Further, the appropriateness of vocabulary both in terms of the age levels to be tested and general familiarity to students was reexamined. Art work was reviewed to insure that content was consistent with the narrative, drawings were accurate and to scale, and labeling was adequate.

Item and test format as a whole were also reexamined at this time. Such considerations as sufficient space for student responses, standard size and terms for unknowns, and possible confusion between labels for an item and information within the item itself were checked. All items were also worked once again as a final verification of expected responses.

From the final pool of 39 items, one item was chosen for the sample item (see Figure 1). It was decided that three of the most difficult items should be administered to 17-year-olds only; three of the easiest items (for 17-year-olds) replaced them for 9-, 11-, and 13-year-olds. Thus, there were 38 items total and 35 items available for each of

these age groups. Separate group-administered test batteries were then prepared for 17-year-olds and for 9-, 11-, and 13-year-olds. Separate batteries were necessary because the items for the 17-year-olds included the stem and questions for all four levels of reasoning whereas the tests for the younger students did not include the extended abstract question.

The two batteries were further organized in two booklets, Booklets 1 and 2, to accommodate most conveniently the two formats in which the items would be administered. Booklet 1 contained items in the basic superitem format. Five test forms of seven items each were created for each age group by randomly assigning items, with the restriction that each content category (except unfamiliar) be represented at least once but no more than twice per form. The assignment was adjusted so that items in the same content category were not contiguous within each form. Booklet 2 contained the same randomly selected 10 items for all ages. The items contained the stem and a question at a single level of reasoning or the stem and two questions in one of the three possible pairwise combinations of levels of reasoning. That is, for 17-year-olds the items contained the stem and level(s) M, R, E, MR, ME, or RE; level U was not included in Booklet 2 for this age group although it was administered in Booklet 1. Using levels U, M, and R, similar items were constructed for 9-, 11-, and 13-year-olds.[1]

---

[1] In addition, an attitude questionnaire, a short verbal scale, and the NAEP student questionnaire were included in each battery.

Stage 5

- The tests were administered during the week of September 14-18, 1981. A central Wisconsin school district serving a community of 32,000 and the surrounding rural area agreed to provide a sample of approximately 300 students in each age group for the administration of the batteries. The school district administrators were extremely cooperative in making arrangements for the testing, particularly in establishing a positive attitude toward the testing among students and parents. A letter publicizing the testing and encouraging full support was sent by direct mail to every parent. After reductions due to absences, underage/overage students, and a few cases of unusable data, the final sample sizes were:

Age	Number
17	303
13	490
11	370
9	308

The test packets containing the two booklets were randomly distributed to students. At the high school, R&D Center staff members assisted by school staff administered both booklets during the first three class periods of one school day with the students assembled in several large group areas. There were two one-hour sittings with a short break between sittings. The mathematics teachers in the middle school administered the tests during math class times on three consecutive days. In this case, both questionnaires and the verbal scale were given the first day followed by the actual tests on the second and third days. At the elementary schools, the two booklets were administered in two one-hour

sittings on consecutive days by classroom teachers or by the building principal.

Finally, the validity of the responses generated in the group-administered test setting was examined about six weeks after the initial administration by means of individual clinical interviews conducted with 12 students at each level by members of the project staff. Each student was administered two superitems. The students were selected at each age level on the basis of the cluster analyses for two of the test forms.

Stage 6

The notion of construct validity implies that the scores on a test can be meaningfully interpreted in terms of related concepts from a psychological theory. By specifying some of the rules of correspondence which connect the theory and data and examining whether or not the data satisfy the theory, one can establish construct validity. In this study we posed three primary and three secondary questions related to the superitems.

Question 1. For each item is the pattern of responses for any student a Guttman true-type response?

The structure of the SOLO taxonomy assumes a latent hierarchical and cumulative cognitive dimension. Consequently, the response structure associated with any level of reasoning determines the response structure associated with all lower levels in the sense that the presence of one response structure implies the presence of all lower response structures. Such response patterns are called Guttman true types (Guttman, 1941). Any deviation from a true type is classified as an error. Then measures

of the extent to which the observed response patterns belong to Guttman true types were used to answer this first question. Three indices were calculated: a coefficient of reproducibility (r), Proctor's (1970) probability of misclassification (p), and an overall chi square. The last value was found by summing the chi square values for all pattern differences between predicted and observed frequencies of patterns of response.

For each index a different criteria was used to determine if a superitem was satisfactory ($r \geq .85$, $p \leq .5$, χ^2 was significant). For 17-year-olds, only 4 superitems had practical problems which indicate they do not reasonably reflect the SOLO taxonomy; 2 superitems were questionable; and 29 were satisfactory. For the 13-year-olds, there were 27 satisfactory superitems; 3 that were questionable; and 5 did not reflect the SOLO taxonomy. For the 11-year-olds, there were 26 satisfactory superitems; 4 questionable superitems; and 5 which did not reflect the SOLO levels. And, for the 9-year-olds, 27 items were considered satisfactory; 3 questionable; and 5 unsatisfactory.

For the 32 superitems that were administered for all four age groups, 20 were satisfactory for all ages. Furthermore, each of the items found questionable or unsatisfactory across all ages appears to have a content validity problem.

In general, this is strong evidence that the superitem format in which items are constructed to fit the SOLO taxonomy forms a Guttman scale. Hence, the results at each age level are consistent with the notion that there are latent cognitive levels which underlie the SOLO taxonomy and that performance is cumulative and hierarchical.

Question 2. From their responses, can the students at each age level be grouped into interpretable groups which reflect the SOLO levels?

The aggregated scores of students on superitems corresponding to the four levels of reasoning in the SOLO taxonomy provide a basis for a possible natural arrangement of subjects into homogeneous groups. If a student's responses to a set of superitems are all Guttman true-type responses, and if the student is at a particular base stage of development, one would expect the average response pattern across several superitems to reflect that base stage of development. The maximum hierarchical clustering method (Johnson, 1970) was used to partition the students on each form and across forms into homogeneous groups based on their score vectors. For the 17-year-olds, the four componnets of the vectors were the aggregated scores on the four taxonomic levels of reasoning: uni-structural (U), multi-structural (M), relational (R), and extended abstract (E); for the younger students there were three components corresponding to the first three levels of reasoning.

Separate cluster analyses were done on the student profiles by form and for a sample across forms at each age level. For the latter analysis seven interpretable groups were identified for the 17-year-old sample. Of this sample 54% are in the M to R range, 31% above R, and 16% below M.

For the 13-year-old sample across forms, eight interpretable groups were formed. The largest single group (50%) were at the M level with another 28% just above or just below level M.

For the 11-year-olds, the cluster analysis of the sample group profiles yield seven interpretable groups with 58% in transition from

U to M. And for the 9-year-olds, the cluster analysis of the sample profiles yielded six groups with 54% of the population around level U.

In all, the interpretability of the cluster profiles across forms indicates the stable influence of cognitive levels of development in the formation of the clusters. Furthermore, the clusters strongly support the utility of the SOLO response categories over the developmental base stages. Clearly, answering the questions in these superitems involves more than level of cognitive development.

Question 3. Does the superitem test format have an effect on the responses to questions at various levels?

It has been assumed that the individual questions within a superitem are not independent. In fact, it is the lack of independence that led Cureton (1965) to his discussion of such superitems.

To measure this relationship, Booklet 2 in each battery of tests consisted of subsets of questions from the total set of superitems. From scores on these tests, it was possible to determine if the questions had an effect on each other by analysis of variance. We assumed that answering a lower level question would facilitate answering a higher level question correctly, but being asked to answer a higher level question would debilitate answering a lower level question correctly.

For 17-year-olds, for the one-way ANOVA for differences of means on the M, R, and E scales when imbedded in different forms, significant differences between means were found in each case. For 13-, 11-, and 9-year-olds, the one-way ANOVA for differences in means on the U, M, and R scales yielded significant differences in each case.

Thus, for all four age groups, the questions within a superitem cannot be considered independent. Furthermore, the results suggest

that asking a lower level question prior to a higher level question increases performance on the latter question, but asking a higher level question decreases performance on a lower level question.

In addition, since Booklet 2 was always given after Booklet 1, the effect of sequence was also to be examined via two additional analyses of variance. The second ANOVA compared means for each reasoning level for independent groups of students on Booklets 1 and 2. The third ANOVA compared the difference scores for students who had the same level of questions in both Booklets 1 and 2. We assumed means for the higher level questions in Booklet 2 would be higher than they were for Booklet 1.

For the 17-year-olds, the two subsequent analyses of variance to examine sequence effects found the differences in means for independent groups and for dependent groups significant on the M and E scales but not on the R scales. Booklet I means were higher on the M scale and Booklet 2 means higher on the E scale.

Similarly, for the 13- and 11-year-olds, significant differences were found for each scale for both independent and dependent groups. Furthermore, the means in both cases were in the expected order.

Finally for 9-year-olds, significant differences were found for the U and R scales but not for the M scale. Also, for both the U and R scales, the means were in the predicted order.

Thus, a sequence effect is apparent. Responding to higher level questions goes up on the second administration of such questions, while responding to lower level questions goes down.

Question 4. What is the reliability of a test made up of superitems?

Since the results of answering Question 3 indicate the questions have an effect upon one another, then the standard procedures for estimating the reliability of a test form are not appropriate. The unit for estimating the reliability is not to be the individual questions but rather the superitems. The internal consistency of a superitem test can be estimated by KR-20 as suggested by Cureton (1965) to counter the effect of correlated errors of measurement produced by the differences among subjects, in general comprehension of the item stem.

The estimated reliability coefficients for the 17-year-olds on the forms and superitems with four questions each ranged from .55 to .82. The estimates for the other three age groups on the forms with superitems with three questions each ranged from .35 to .75.

These coefficients are not high but are considered reasonable since each form only contained seven superitems and there was little variability on the lower or upper level questions in some populations.

Question 5. What is the reading level of each superitem?

Since we planned to administer the same superitems to students of ages 9, 11, 13, and 17, it was reasonable to check on the reading level of the textual information in the superitems. After all mathematics terms had been deleted, the text was entered into a textual analysis computer program, and four readability indices found. The Flesch Index (Flesch, 1948) is a predicted score based on average word length (in syllables) and average sentence length (in words). The Dale Index

(Dale, 1948) is a predicted score based on average sentence length and number of unfamiliar words (words not in the Dale list of 3000 words). FOG Index (Gunning, 1951) is based on average sentence length and number of high caliber words (words of three or more syllables). Fry Index (Fry, 1967-68) is based on average number of sentences and the average number of syllables.

For the 17-year-old population, these indices were based on the total superitem of stem and four questions; the 17-year-olds who were to answer E questions needed to understand some new information in those questions. For the other populations, these indices were based only on the stem and U question; it was felt that the stem and U question contained the basic information which needed to be read and understood. The overall results of the readability analysis for 17-year-olds indicated that all superitem stems and questions were of reading difficulty appropriate to twelfth graders. For 13-year-olds, four superitems were judged to be inappropriately difficult for them and several more superitems were marginal; overall the superitems seemed appropriate for students at this age. For 11-year-olds, the readability of test items is questionable, 12 of 35 items were too difficult and several were marginal. Finally, for 9-year-olds, 24 items were judged too difficult, and several were marginal.

Hence, the reading difficulty of the problem-solving test in its present format does not seem appropriate for 9-year-olds. It is marginally appropriate for 11-year-olds, and it is adequate for both 13-year-olds, and 17-year-olds.

Question 6. What is the relationship of a student's pattern of responses on a group-administered superitem test with his/her pattern to similar items given in an interview situation?

Under the assumption that valid data are gathered in individual interview situations, but because of cost it would be useful to gather data via a group administration, we decided to see if the patterns of responses differed in the two situations. In fact, we assumed the interview scores would be slightly higher because reading or procedural errors can be corrected, but the patterns of responses should reflect the same underlying base stage of development.

The interview data were gathered on a very small sample of students, twelve at each age level, and each student was asked to respond to two superitems. Understanding these limitations for the comparisons, performance on the interview questions was higher than on group-administered questions. Several reasons for the differences were apparent. For U and M questions, the interviewers noted several instances where students raised questions which clarified their understanding of questions or got them to correct a procedure error. For R and E questions, prompts or answers to questions (or lack of answers) caused students to rethink the question. And for the 9-year-olds, since the questions were read to the students in the interview situation, readability was not a source of error.

Nevertheless, the overall pattern of responses continue to strongly support the SOLO taxonomy. What it indicates is the group-administered testing situation adds another factor to the response level interpretation.

Stage 7

To complete the project, three reports have been prepared. The first is:

Romberg, T. A., Collis, K. F., Donovan, B. F., Buchanan, A. E., and Romberg, M. N. The development of mathematical problem-solving superitems.

In this report the details of what we did in Stages 1-4 summarized above are reported.

The second report is:

Romberg, T. A., Jurdak, M., Collis, K. F., & Buchanan, A. E. Construct validity of a set of mathematical superitems.

The extensive data collection and analyses related to the six questions in Stages 5 and 6 summarized above are reported in this document.

The third report is this document. Attached for NIE and ECS, but not for general distribution, are two appendices. The first is a set of test booklets as administered in this study. The second is the complete set of items with technical details and comments about each item. The superitems developed for this study are the property of the Education Commission of the States; they are not available for research or general use.

Finally, other data were gathered in this study (an attitude questionnaire, general background information on both students and the schools, and a verbal ability scale). We anticipated that if we were successful in developing the superitems and demonstrating their construct validity funds would be forthcoming to carry out a secondary analysis relating this additional data to these results. At this time funds are not available for this analysis.

Conclusions

The purpose of this project was to develop a set of superitems which reflected the levels of reasoning posited in the SOLO taxonomy, to validate those items, and to estimate the utility of the items for large scale assessment.

With care and some difficulty, the staff was able to construct a content valid set of items and prepare them for administration. Analysis of the data gathered revealed the following about the superitems. First, the majority of items were Guttman true-type items with response patterns matching the assumed latent hierarchical and cumulative cognitive dimension. Second, from the question profiles for each student, clusters of students were formed and the profiles for those clusters were interpreted in terms of developmental base stages. Together these findings gave strong support to the validity of the sequence of SOLO levels. And third, the utility of the SOLO approach to superitem construction and interpretation of responses is also apparent. Answering content-based questions at varying levels requires more than level of cognitive development. Thus, the SOLO interpretation of responses is more useful for educators and researchers in describing level of reasoning on school related tasks.

Recommendations for Future Use of the Items

Our intent was to develop a set of items which could be used in large scale assessment projects like the National Assessment of Educational Progress. We believe we have been successful in this effort. Thus, we make the following recommendations based on our experience.

1. Superitems should be individually examined and selected for use.

   This was a superitem development project. The adequate superitems reflect different content areas, and have different readability levels. Thus, we do not recommend use of the booklets developed in this study in their present form. The superitems in the booklets reflect our considerations about how to establish construct validity.

2. Initial selection of superitems should be based on content area of interest.

   The SOLO taxonomy assumes that the response level of any student to such superitems depends on base developmental level and other factors such as familiarity of content. Thus, if a researcher is interested in level of reasoning for a group of students, the superitems should be selected to take many other factors into account. In particular, the content of the items should be considered first.

3. The number of superitems selected for use should depend on the unit of investigation.

   The superitems were developed for large-scale group administration. However, they could also be used for diagnostic purposes for individuals or for research purposes with small groups. A superitem with four questions (one each for levels U, M, R, and E) takes an average of seven minutes for a typical 17-year-old, and one with three questions (U, M, and R) takes an average of 5 minutes for typical 13- or 11-year-olds and some 9-year-olds. With this in mind, for large scale administration where the unit of investigation

is a population (a class, a school, a district, or even an age level group like those tested by NAEP), one or two superitems per booklet would be sufficient for any one content area. Researchers interested in forming clusters of students for a particular study should select about 10 superitems. This would increase the reliability of the form and increase the probability of forming interpretative groups. And finally, for the scholar interested in interviewing a small group of students, three to five superitems should be sufficient to establish the level of reasoning for any student.

4. Group administration of these superitems to students 9 years old or younger is not recommended without further study. Although useful information was derived for 9-year-old children, the low indices of readability make any interpretation suspect. If the items are read to students, however, then we see no problem.

References

Collis, K. F., & Biggs, J. B. Classroom examples of cognitive development phenomena: The SOLO taxonomy. Report prepared at conclusion of an Educational Research and Development Committee funded project, University of Tasmania, 1979.

Cureton, E. E. Reliability and validity: Basic assumptions and experimental designs. Educational and Psychological Measurement, 1965, 25, 326-346.

Dale, E. A formula for predicting readability. Educational Research Bulletin, 1948, 27, 11-20.

Flesch, R. A new readability yardstick. Journal of Applied Psychology, 1948, 32, 221.

Fry, E. A readability formula that saves time. Journal of Reading, 1967-68, 11, 513.

Gunning, R. The techniques of clear writing. New York: McGraw-Hill Book Company, 1952.

Guttman, L. The quantification of a class of attributes: A theory and method for scale construction. In P. Horst (Ed.), The prediction of personal adjustment. New York: Social Science Research Council, 1941, 319-348.

Johnson, S. C. Hierarchical clustering schemes. Psychometrika, 1970, 35, 241-254.

Proctor, C. H. A probablistic formulation and statistical analysis of Guttman scaling. Psychometricka, 1970, 35, 73-78.

Wearne, D. C., & Romberg, T. A. DMP Accountability Tests (Working Paper No. 217). Madison: Wisconsin Research and Development Center, 1977.